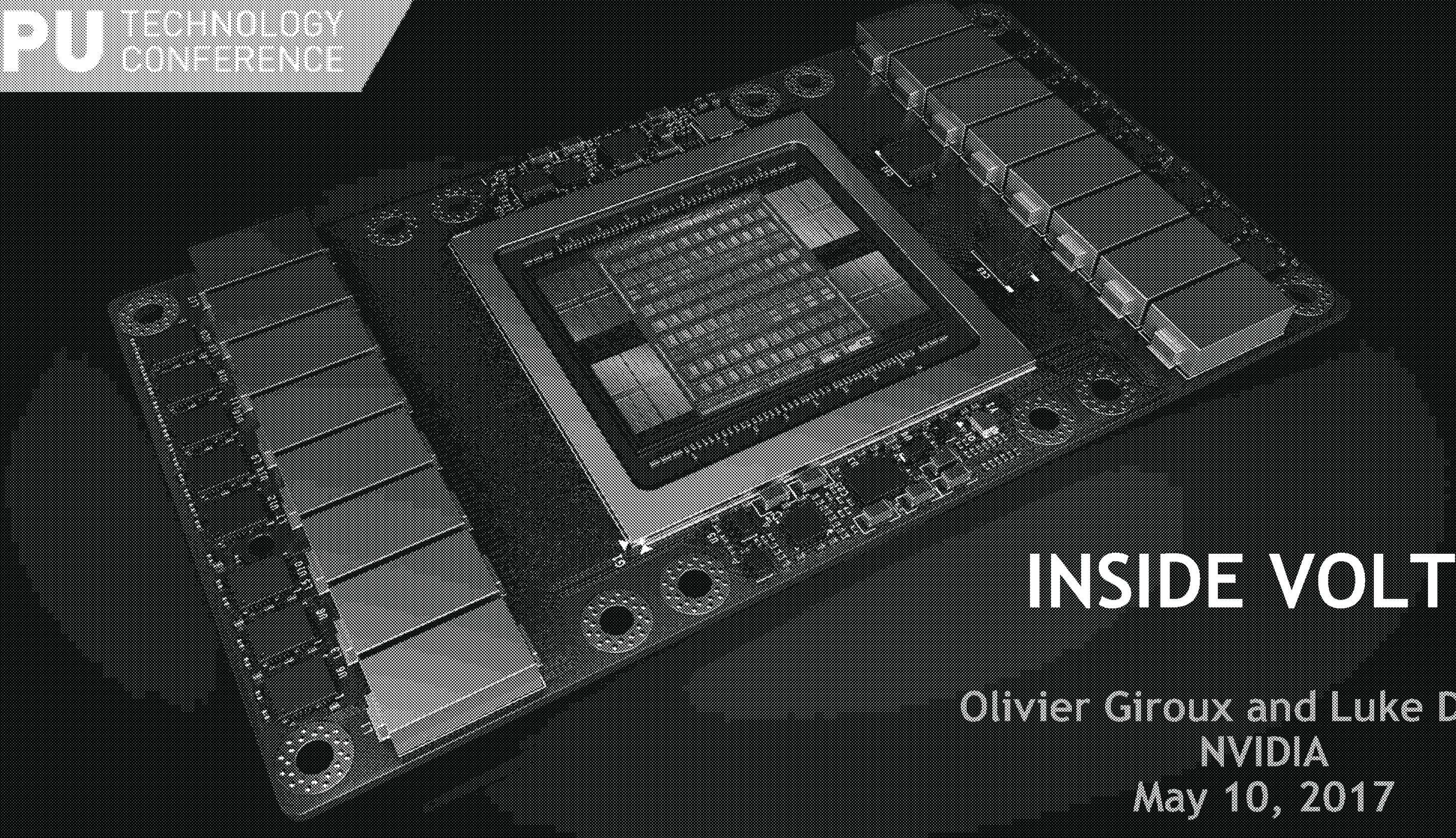
Exhibit 3

GPU TECHNOLOGY CONFERENCE



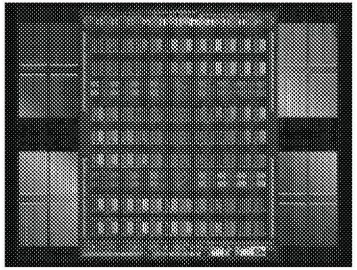
INSIDE VOLTA

Olivier Giroux and Luke Durant
NVIDIA

May 10, 2017

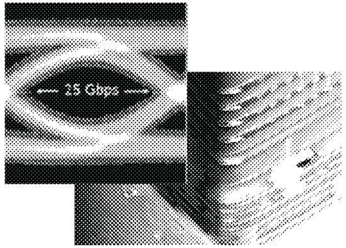
INTRODUCING TESLA V100

Volta Architecture



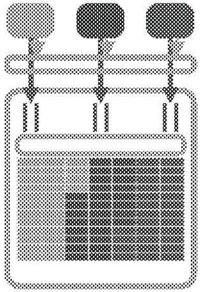
Most Productive GPU

Improved NVLink & HBM2



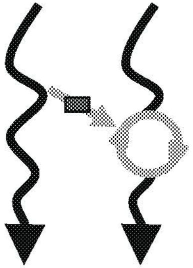
Efficient Bandwidth

Volta MPS



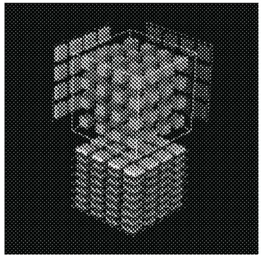
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



120 Programmable TFLOPS Deep Learning

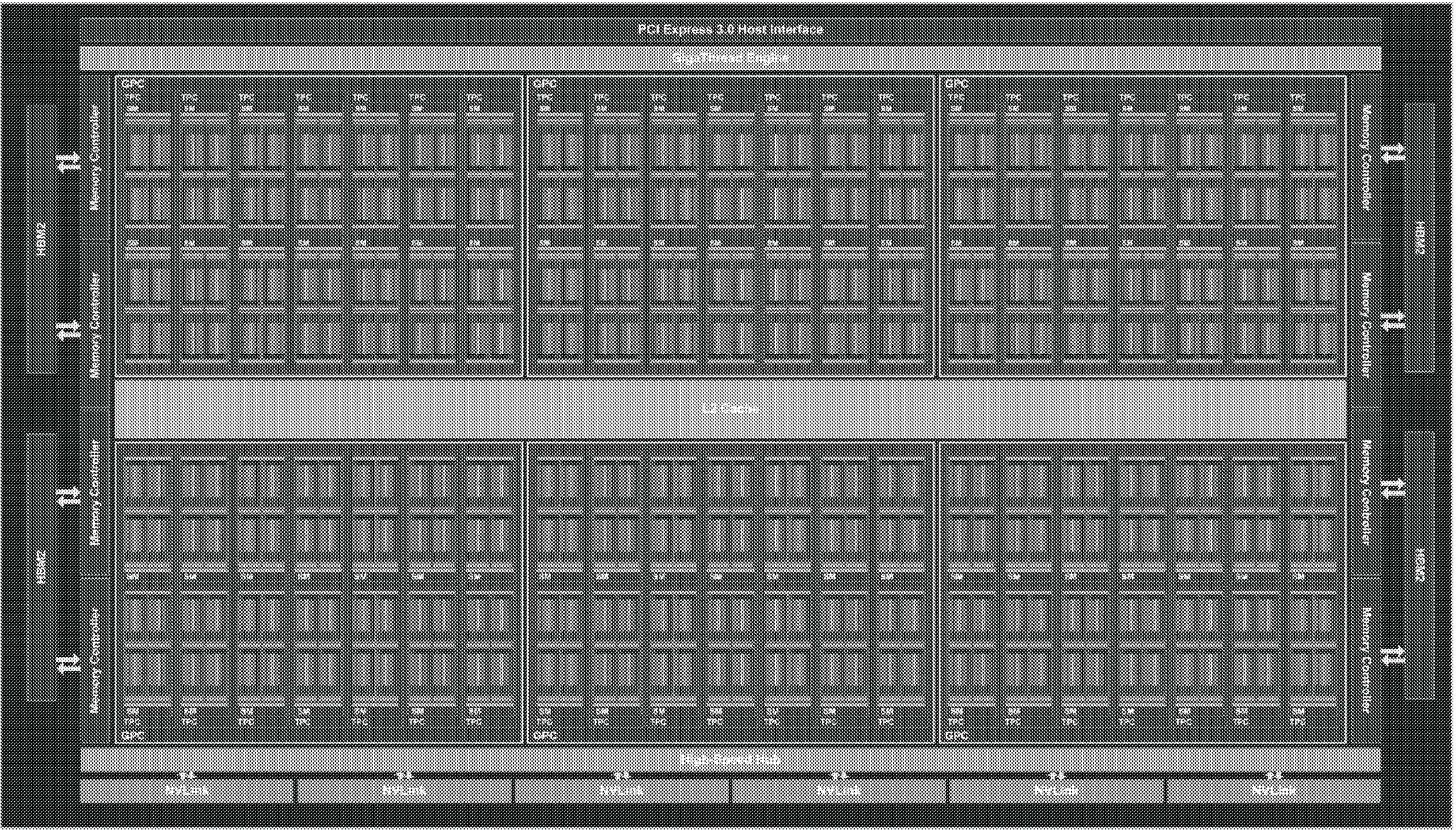
The Fastest and Most Productive GPU for Deep Learning and HPC

TESLA V100

21B transistors
815 mm^2

80 SM
5120 CUDA Cores
640 Tensor Cores

16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

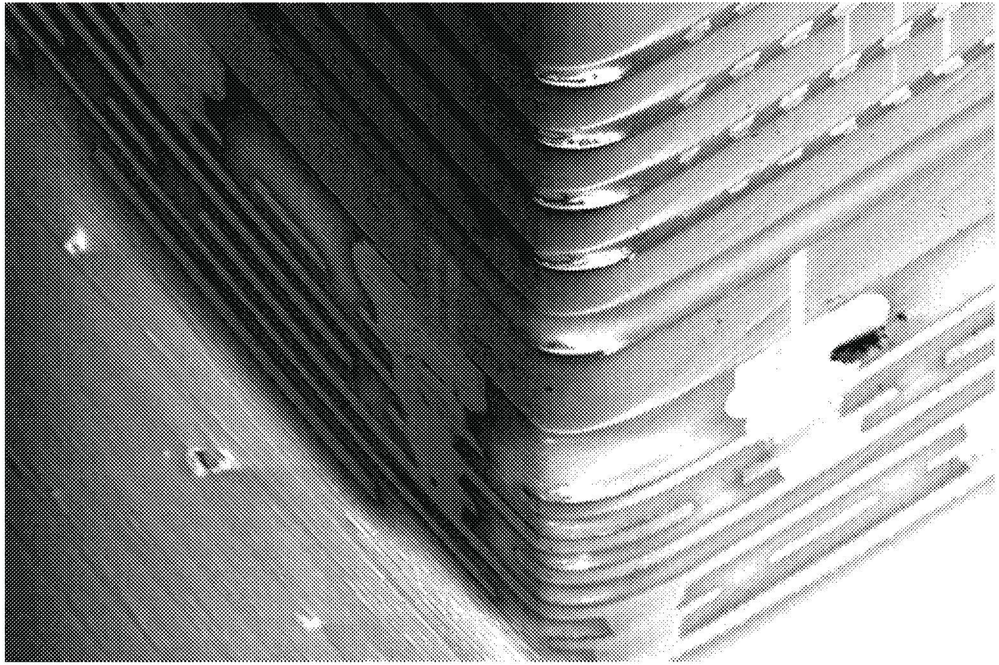


*full GV100 chip contains 84 SMs

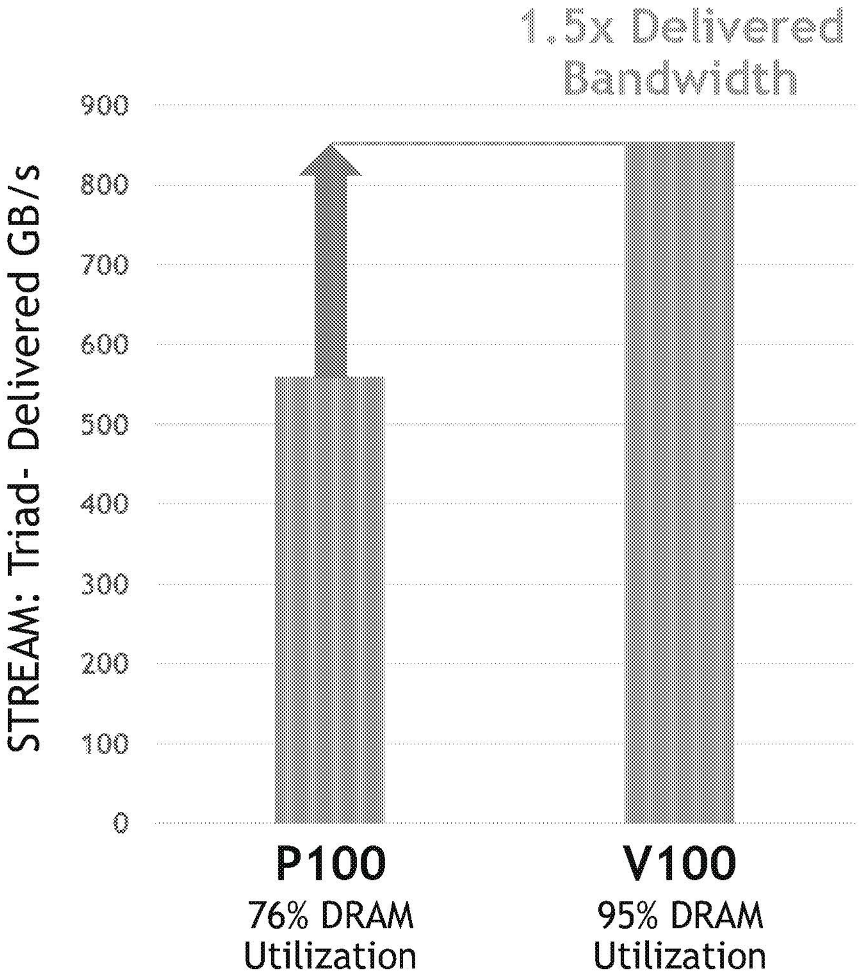
GPU PERFORMANCE COMPARISON

	P100	V100	Ratio
Training acceleration	10 TOPS	120 TOPS	**12x**
Inference acceleration	21 TFLOPS	120 TOPS	**6x**
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	**1.5x**
HBM2 Bandwidth	720 GB/s	900 GB/s	**1.2x**
NVLink Bandwidth	160 GB/s	300 GB/s	**1.9x**
L2 Cache	4 MB	6 MB	**1.5x**
L1 Caches	1.3 MB	10 MB	**7.7x**

NEW HBM2 MEMORY ARCHITECTURE



HBM2 stack



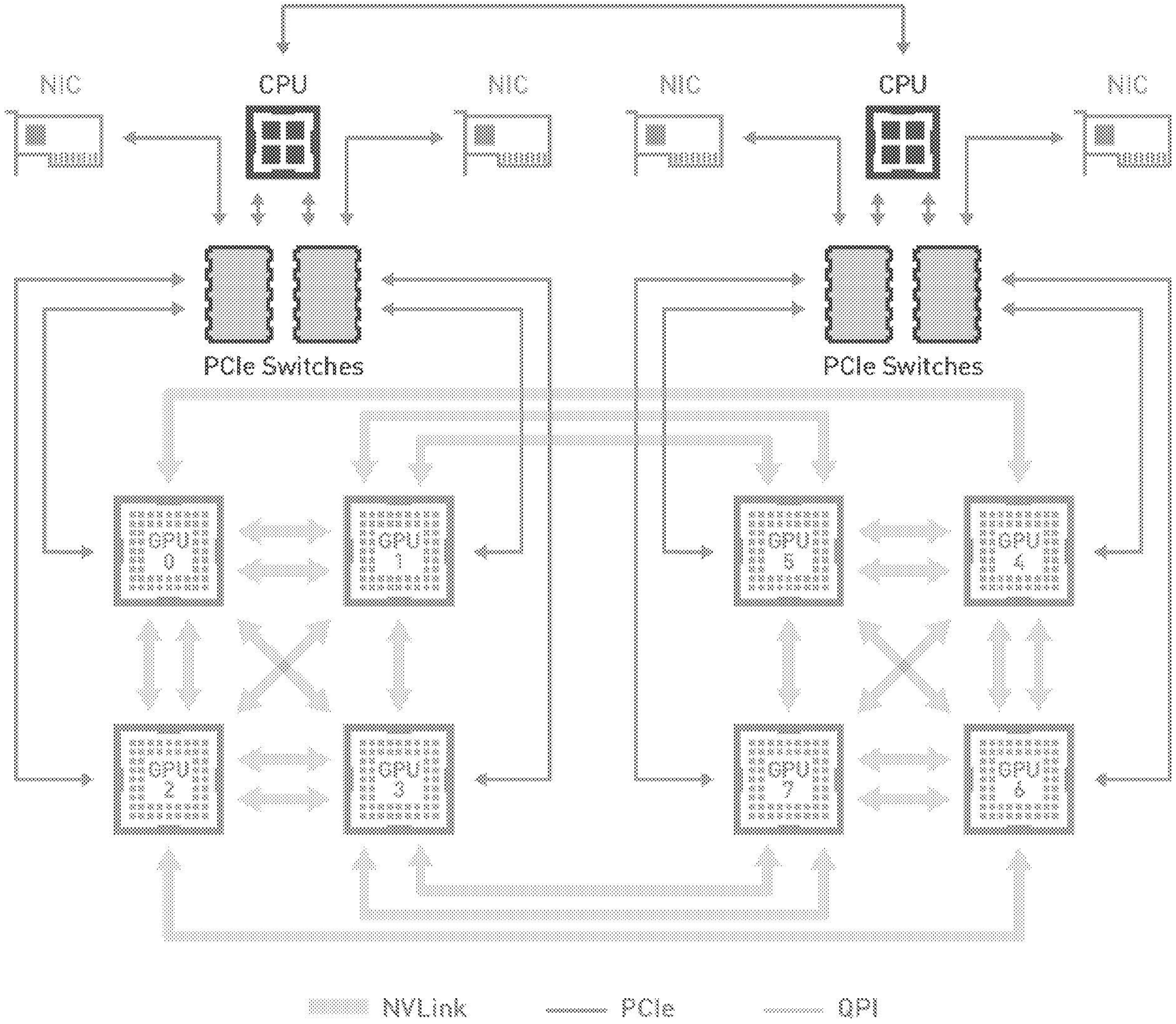
V100 measured on pre-production hardware.

VOLTA NVLINK

300GB/sec

50% more links

28% faster signaling


NIC
CPU
NIC
NIC
CPU
NIC
PCIe Switches
PCIe Switches
GPU 0
GPU 1
GPU 5
GPU 4
GPU 2
GPU 3
GPU 7
GPU 6
NVLink
PCIe
QPI